

Multi-scale Population and Mobility Estimation with Geo-tagged Tweets

Jiajun Liu, Kun Zhao, Saeed Khan, Mark Cameron, Raja Jurdak
Commonwealth Scientific and Industrial Research Organisation (CSIRO), Australia
{firstname.lastname}@csiro.au

Abstract—Recent outbreaks of Ebola and Dengue viruses have again elevated the significance of the capability to quickly predict disease spread in an emergent situation. However, existing approaches usually rely heavily on the time-consuming census processes, or the privacy-sensitive call logs, leading to their unresponsive nature when facing the abruptly changing dynamics in the event of an outbreak. In this paper we study the feasibility of using large-scale Twitter data as a proxy of human mobility to model and predict disease spread. We report that for Australia, Twitter users’ distribution correlates well the census-based population distribution, and that the Twitter users’ travel patterns appear to loosely follow the gravity law at multiple scales of geographic distances, i.e. national level, state level and metropolitan level. The radiation model is also evaluated on this dataset though it has shown inferior fitness as a result of Australia’s sparse population and large landmass. The outcomes of the study form the cornerstones for future work towards a model-based, responsive prediction method from Twitter data for disease spread.

I. INTRODUCTION

Information on human mobility is essential for modeling and predicting disease spread. This information can be obtained using traditional census data [1], or new sensing technologies such as mobile phones [2], [3], RFID [4] and WIFI logging [5]. For example, census data of commuting flows [1] has been used in the simulation of multi-scale human mobility and spatial spread of infectious diseases. Traffic patterns in air transportation networks have been used to model human mobility in response to a disease outbreak [6]. The study in [3] reports that human mobility can be inferred by the frequency of mobile phone calls between two locations and their geographical distance. Mobile phone records coupled with malaria prevalence data have been used to identify the dynamics of human carriers of the parasite responsible for spreading disease across regions in Kenya [7]. Numerous methods of using mobile phone records as a proxy for sensing human mobility are surveyed in [2]. More recently, advances of sensory and information technologies have also enabled us to study disease spread at a fine-grained level by using RFID [4] or WIFI logging [5] to obtain high-resolution human contact patterns in a limited range. However, due to privacy concerns for call records and the tremendous human effort involved for censuses, these data sources are not always accessible, and they suffer from a range of limitations, such as the latency of census data, the low spatiotemporal resolution of mobile phone records, and the limited sensing range of RFID and WIFI logging. The acquisition of real-time,

high spatiotemporal resolution and large-scale human mobility patterns for fine-grained modeling and predicting epidemic dynamics remains a challenge.

Twitter appear to be a promising proxy to overcome this challenge. For example, geo-tagged tweets can provide high-precision location information, and in the meanwhile offer comprehensive metadata with higher granularity and resolution. Moreover, tweets are generated continuously in large volume and are generally available to public, which provides timely and accessible information on human mobility. Although twitter constitutes a valuable data source for tracking human mobility, its potential power in modelling human mobility and epidemic spreading has not been fully exploited.

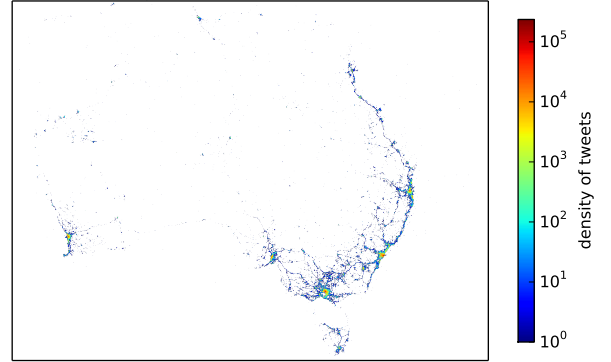


Fig. 1. The visualization of the geo-tagged Tweets highlights Australia’s most dense areas and roughly resembles its population distribution.

In this paper, we report some preliminary results for the estimation of population distribution and human mobility flows in Australia using a large-scale dataset of 6,304,176 geo-tagged Tweets generated by 473,956 unique users. We find that the population distribution can be roughly estimated from geo-tagged Tweets. We apply two seminal models, namely gravity model and radiation model, to estimate mobility flows. In contrast to the conclusion that the radiation model performs better than the gravity model [8], the gravity model demonstrates superior goodness of fit as well as better robustness to changes in geographic scales in the estimation of mobility flows in Australia. The likely ingredients of such results include the unique characteristics of Australia’s geographic features and its sparse population distribution.

TABLE I
STATISTICS OF THE DATASET

Range of longitude	Range of latitude	Collection period	No.Tweets	No.unique users	Avg.Tweets /user	Avg.waiting time	Avg.no. locations/user
[112.921112, 159.278717]	[-54.640301, -9.228820]	Sept.2013-Apr.2014	6,304,176	473,956	13.3	35.5hr	4.76

II. GEO-TAGGED TWEETS

Though affected by the inevitable sampling bias as well as by the individual tweeting dynamics, geo-tagged Tweets are generally considered a high quality data source for capturing human mobility [9], [10]. Particularly at the national or the global scale, the collective tweeting efforts are believed to reveal the “true” mobility patterns to a great extent [9]. In addition, geo-tagged Tweets exhibit a few favourable properties that elude other data sources, such as its massive user base, near-instantaneous updates, and public availability, all of which are crucial for modeling an emergent disease outbreak.

In light of these advantages, we collected all Tweets sent from Australia from September 2013 to April 2014. We show the distribution of the number of Tweets per user and the distribution of the waiting time interval between consecutive Tweets to further explore the tweeting dynamics in Figure 2. Both the distribution of the number of Tweets per user and the distribution of waiting time span at least eight decades and exhibit a heavy tail. Evidently, similar to other countries [9], the tweeting behaviors of the Australian population also exhibit the Pareto principle. The distribution of the number of Tweets per user essentially follows a power-law distribution whereas the distribution of time-intervals demonstrates substantial heterogeneity which indicates that the tweeting dynamics are likely to be influenced by multiple factors, such as prioritized task handling in human behaviors [11].

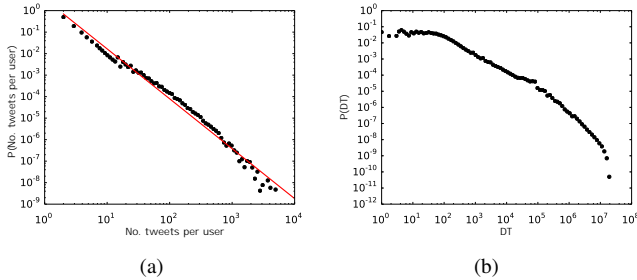


Fig. 2. Both the number of Tweets per user and the time intervals between consecutive Tweets are heavy-tailed.

In Table I we provide more basic statistics of the collected Tweets. We use the longitude and latitude ranges to filter the Tweets of interest published from Australia. Over the 7 months collection period, the average number of Tweets per user is only 13.3, however a small portion of the Twitter population demonstrates enthusiasm in sharing their daily life, with the numbers of users with more than 50, 100, 500, 1000 Tweets being 23462, 10031, 766 and 180 respectively.

III. POPULATION ESTIMATION

To gain more insight into the feasibility of using Twitter data to model and predict the spread of infectious diseases, we

examine the correlation between the geographic distribution of Twitter users and the Australian population distribution reported by census ¹. The study is conducted at three scales, namely national level (20 most populated cities in Australia), state level (20 most populated cities in the state of New South Wales), and metropolitan level (20 most populated suburbs in Sydney) to verify the robustness to geographic scales. A search radius of 50km, 25km and 2km is used for the three scales respectively when extracting number of Tweets, number of users and mobility from Tweets. For the three scales, the average distances between areas are 1422km, 341km and 7.5 km respectively.

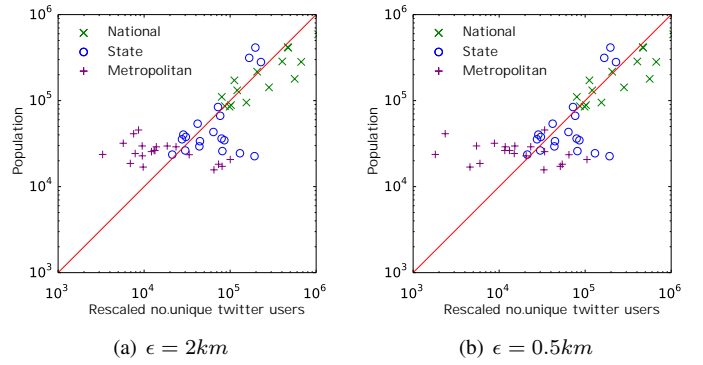


Fig. 3. The rescaled “Twitter population” ($Cp^{Twitter} \approx p^{Census}$, where C is a rescaling factor) of the examined 60 areas at three different geographic scales shows strong correlation to the population distributions in the real-world. ϵ is the search radius to derive Twitter-based population.

Figure 3(a) depicts the correlation between the number of unique Twitter users and the census-based population of the examined areas at the three scales. In total, 60 samples are shown in the figure, with 20 points for each geographic scale. Overall we have a Pearson correlation coefficient of 0.816 between the “Twitter population distribution” and the real-world population distribution, with a two-tailed p-value of 2.06×10^{-15} , indicating a strong correlation between the two. We argue that based on this result, estimating the population distribution from the geo-tagged Tweets is feasible.

In addition, the correlation appears to weaken as the population size and geographic scale decrease. That is, the estimation for National aligns best with the census data, and for State a few outliers begin to manifest, while for Metropolitan the points scatter further more (though still centered at $y = x$). It could be that the sample size of Tweets tend to decrease for smaller geographic areas, making the sample bias a more significant factor. However, as we extract the median numbers of users for the three scales, we find that for National, State and

¹<http://www.abs.gov.au/AUSSTATS/abs@.nsf/DetailsPage/3218.02012-13>

Metropolitan, the figures are 4166, 743, and 3988 respectively. We observe that despite a smaller sample size, State yields better prediction power than Metropolitan, which indicates that the sample size has limited influence on the correlation. Now considering that the noises introduced while conducting censuses and while extracting the Tweets and users tend to have greater effect on smaller areas, in attempt to investigate such an effect, we change the search radius of Metropolitan to 0.5km and generate Figure 3(b). This results in significant increase of error for Metropolitan and hence we argue that the sensitivity to the edges of the areas and search radius is likely to be a prominent factor to the differences of the correlation for different scales. A more in-depth discussion around this phenomenon will be provided in future work.

IV. MOBILITY ESTIMATION

To explore the feasibility of mobility estimation with Twitter data, first we extract the mobility from Tweets by counting how many pairs of consecutive Tweets appear first at the source area and then the destination area, and then we capture the mobility between areas using two models, namely the Gravity model and the Radiation model:

Gravity: Similar to Newton's law of gravitation, the Gravity model suggests that mobility between two places, namely an origin and a destination is proportional to the product of populations of these two places, and is inversely proportional to the power law of distance between them [12]. It is considered one of the fundamental models to predict not just human mobility, but also trade flows between countries as well as communications volume between cities. However, it has also demonstrated some limitations. It is considered to work well for transitions over shorter distances, but for larger distances the results are sometimes inconsistent, thus putting its universality in question.

Radiation: The Radiation model follows the analogy of the particle diffusion model in which particles emitted at a source have certain probability of being absorbed by a destination location. According to this model [8], the absorption probability depends on the origin's population and the destination's population. In addition, it is also determined by the population within a circle centered at the origin, with the radius equal to the distance between the origin and the destination (excluding the origin and the destination themselves).

In our scenario, they are formally defined as:

a) *Gravity Model (4 Parameters):*

$$P \propto C \frac{m^\alpha n^\beta}{d^\gamma} \quad (1)$$

b) *Gravity Model (2 Parameters):*

$$P \propto C \frac{mn}{d^\gamma} \quad (2)$$

c) *Radiation Model:*

$$P \propto C \frac{mn}{(m+s)(m+n+s)} \quad (3)$$

where α , β , and γ are the models' fitting parameters, C is the scaling parameter. m and n represent the population of

the source and the destination respectively. d is the distance between the two areas. s is defined as the total population within radius d from the center of the source area with the source and destination areas excluded. For Gravity models, given a series of m , n and d values, the parameters α , β , and γ can be estimated from least-square fitting after taking logarithm of the formulas. In the rest of the paper we refer to the three models as Gravity 4Param, Gravity 2Param, and Radiation.

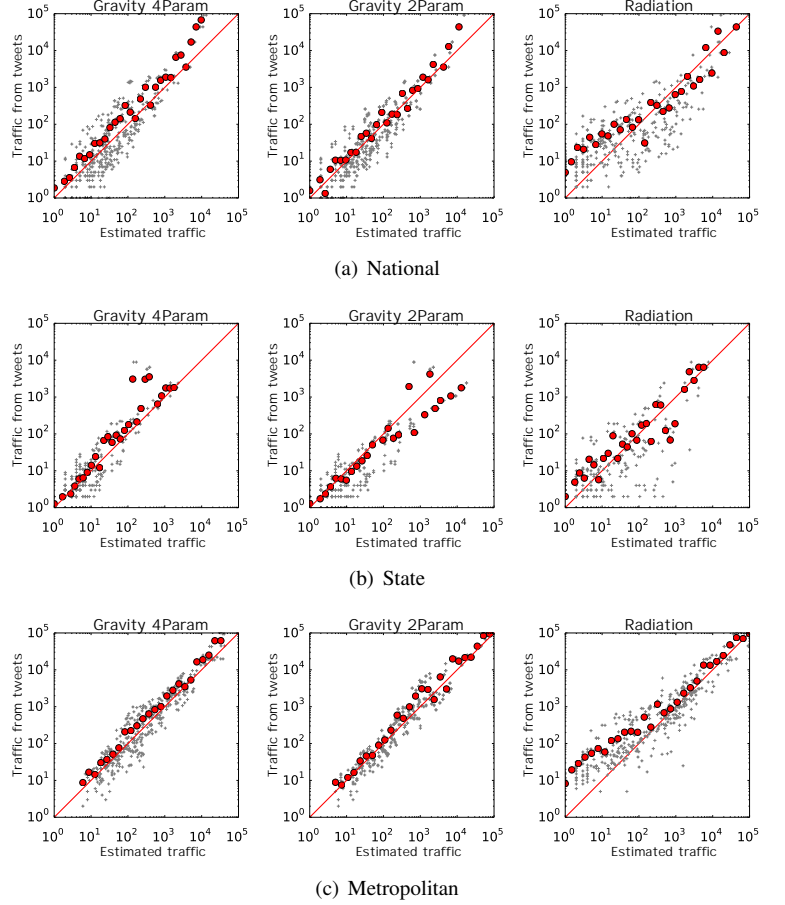


Fig. 4. Mobility estimation performance of Gravity 4Param, Gravity 2Param and Radiation at three scales. Overall Gravity demonstrates advantages over Radiation. Radiation's results are likely due to the fact that Australia's sparse and uneven population distribution deviates from Radiation's implication of a continuous dispersion of population from major centers.

The performance of the models is reported in Figure 4. The x -axis is the estimated mobility from the models, the y -axis is the mobility extracted from Tweets, the grey crosses are the pairs of the estimated and extracted mobility, the red dots are the averaged values in the bins after logarithmic binning, and the red line represents $y = x$. A tighter spread of the grey and red markers around the red line indicates better accuracy, whereas dots falling right of the red line suggesting overestimation and left suggesting underestimation.

Visually Gravity models show better performance of capturing the mobility pattern between areas as well as better robustness to various geographic scales over Radiation. For

example, For National (Figure 4(a)), it is shown that for Gravity 2Param, with the exception of a few cases (particularly at larger number scales in the upper right corner), the grey crosses and red dots tightly surround the red line, suggesting that in general the estimated mobility is fairly accurate. A closer examination suggests that the estimation error is roughly bounded by one decade, with most of the cases being smaller. Gravity 4Param shows similar distributions of the grey crosses and the red dots visually, but with slightly greater dispersions. On the other hand, we could see the grey for Radiation scatter loosely around the red line, leading to a maximal error across almost two decades, with moderate tendency to underestimate. For State (Figure 4(b)) we note that Gravity 4Param appears to have the best results, while Gravity 2Param has a slight tendency to overestimate for larger numbers ($> 10^3$). Radiation yields many overestimations and spans its errors in three decades. Figure 4(c) indicates that at this scale both the Gravity models work quite well, while Radiation shows a strong tendency to underestimate for smaller numbers.

The performance is also measured quantitatively with two metrics, i.e. the Pearson correlation coefficient between the estimated mobility and the Twitter mobility, and the HitRate@50% (percentage of estimates which have smaller than 50% relative errors). Table II shows these results. For each cell, the upper number is the Pearson coefficient and the lower the HitRate@50%. For each scale and metric, the highest performance is highlighted. It is evident that, quantitatively, Gravity 2Param provides the best performance overall, which aligns well with the previous conclusions from visual inspection.

The results indicate an interesting finding as they contradict the conclusion from [8] that Radiation captures human mobility better than Gravity. That is, Radiation's advantages are not universal, and they may not suit countries that have sparsely and unevenly distributed population, such as Australia or Canada. Unlike U.S.A. where a large population spreads relatively evenly across the country, Australia's population concentrates heavily along its coastline, creating areas with extremely low population densities between populated areas. This feature renders Radiation's underlying assumption that population density decays more smoothly from dense centers unsuited for Australia.

TABLE II
MODEL PERFORMANCE MEASURED BY THE PEARSON CORRELATION COEFFICIENT (UPPER) AND THE HITRATE@50% (LOWER)

	Gravity 4Param	Gravity 2Param	Radiation
National	0.877 0.330	0.912 0.397	0.840 0.184
State	0.893 0.487	0.896 0.397	0.742 0.166
Metropolitan	0.948 0.53	0.963 0.600	0.918 0.397

As we verify the model performance to estimate the mobility extracted from Tweets, we argue that by replacing m and n with the population from census, it is feasible to estimate the real-world mobility between areas in Australia. We will test this proposal in future work.

V. CONCLUSION

In this paper we explore the feasibility of using geo-tagged Tweets to estimate population distribution and mobility for Australia. We first evaluate the correlation between the estimated population to the census-based population, and then follow with a comparative study modeling the mobility extracted from Tweets with the gravity model and the radiation model. To verify the robustness of population estimation and mobility estimation to different geographic scales, both the experiments are conducted at three scales that represent the national, state and metropolitan levels. We report that 1) it is feasible to estimate population distribution from geo-tagged Tweets, particularly for dense areas 2) for countries that have uneven population distributions such as Australia, Gravity model appears a better model to estimate mobility than Radiation. In future we will further improve the model accuracy by incorporating census data of higher resolutions, evaluate model performances with more metrics and at more varieties of distances scales, and use the models to devise a framework for the prediction of disease spread.

REFERENCES

- [1] D. Balcan, V. Colizza, B. Gonçalves, H. Hu, J. J. Ramasco, and A. Vespignani, "Multiscale mobility networks and the spatial spreading of infectious diseases," *Proceedings of the National Academy of Sciences*, vol. 106, no. 51, pp. 21 484–21 489, 2009. [Online]. Available: <http://www.pnas.org/content/106/51/21484.abstract>
- [2] S. Jiang, G. A. Fiore, Y. Yang, J. Ferreira, Jr., E. Frazzoli, and M. C. González, "A review of urban computing for mobile phone traces: Current methods, challenges and opportunities," in *Proceedings of the 2Nd ACM SIGKDD International Workshop on Urban Computing*, ser. UrbComp '13. New York, NY, USA: ACM, 2013, pp. 2:1–2:9. [Online]. Available: <http://doi.acm.org/10.1145/2505821.2505828>
- [3] V. Palchykov, M. Mitrović, H.-H. Jo, J. Saramki, and R. K. Pan, "Inferring human mobility using communication patterns," *Scientific Reports*, vol. 4, no. 6174, 2014.
- [4] K. Zhao, J. Stehlé, G. Bianconi, and A. Barrat, "Social network dynamics of face-to-face interactions," *Physical Review E*, vol. 83, no. 5, p. 056109, 2011.
- [5] A. Chaintreau, P. Hui, J. Crowcroft, C. Diot, R. Gass, and J. Scott, "Impact of human mobility on opportunistic forwarding algorithms," *Mobile Computing, IEEE Transactions on*, vol. 6, no. 6, pp. 606–620, June 2007.
- [6] S. Meloni, N. Perra, A. Arenas, S. Gómez, Y. Moreno, and A. Vespignani, "Modeling human mobility responses to the large-scale spreading of infectious diseases," *Scientific reports*, vol. 1, 2011.
- [7] A. Wesolowski, N. Eagle, A. J. Tatem, D. L. Smith, A. M. Noor, R. W. Snow, and C. O. Buckee, "Quantifying the impact of human mobility on malaria," *Science*, vol. 338, no. 6104, pp. 267–270, 2012. [Online]. Available: <http://www.sciencemag.org/content/338/6104/267.abstract>
- [8] F. Simini, M. C. González, A. Maritan, and A.-L. Barabási, "A universal model for mobility and migration patterns," *Nature*, vol. 484, pp. 96–100, April 2012.
- [9] B. Hawelka, I. Sitko, E. Beinat, S. Sobolevsky, P. Kazakopoulos, and C. Ratti, "Geo-located twitter as proxy for global mobility patterns," *Cartography and Geographic Information Science*, vol. 41, no. 3, pp. 260–271, 2014.
- [10] M. R. Frank, L. Mitchell, P. S. Dodds, and C. M. Danforth, "Happiness and the patterns of life: A study of geolocated tweets," *Scientific Reports*, vol. 3, no. 2625, 2013.
- [11] A.-L. Barabási, "The origin of bursts and heavy tails in human dynamics," *Nature*, vol. 435, p. 207, 2005.
- [12] G. K. Zipf, "The p1 p2/d hypothesis: On the intercity movement of persons," *American Sociological Review*, vol. 11, no. 6, pp. 677–686, 1946.